Utilizing Non-Redundant Association Rules from Multi-Level Datasets

Gavin Shaw, Yue Xu & Shlomo Geva
Faculty of Information Technology, Queensland University of Technology, Brisbane, Australia
gavin.shaw@.qut.edu.au, yue.xu@qut.edu.au, s.geva@qut.edu.au

Abstract

Association rule mining and recommender systems are two popular methods for obtaining knowledge and information from datasets. However, both of these methods suffer from limitations. Traditionally association rule mining has focused on extracting as many rules as possible from flat datasets. More recently, issues over the number of rules and obtaining rules from datasets with multiple concept levels have come into focus. Recommender systems have been popular with users when it comes to helping find similar interests to those they already have. However, recommender systems suffer from two major problems, cold start and novelty.

The aims of our research is to develop an approach for extracting non-redundant multi-level and cross-level association rules from datasets with multiple concept levels and utilise them in a recommender system with the aim of potentially solving the cold start and novelty problems.

1. Introduction

Association rule mining and recommender systems are popular data mining techniques. Association rule mining is a process for finding associations, relations, patterns etc in large datasets with minimal human effort and has been proven to be a success. This technique has found uses in many fields/areas [14]. However, successful application of the extracted rules into real world problems is often restricted by the quality of said rules. Dealing with the quality of association rules, especially multi-level and cross-level rules has drawn little attention. After deriving a set of association rules it is normal to measure the quality of these rules. This can be difficult and most of the existing measures were designed with association rules from flat datasets in mind. Successful application of association rules to problems is heavily dependant on the completeness, accuracy and quality of the rules. Thus obtaining good association rules, especially from datasets with multiple concept levels (or hierarchy/taxonomy) is an important research area.

Recommendations play a major role in a person's everyday life. With the large amount of information available there is a need for automated recommendation methods. Recommender systems provide methods for finding important and relevant information or knowledge from large datasets or data collections with minimal user involvement. These systems have many practical applications. Despite their advantages recommender systems still have limitations, mainly two problems known as cold start and novelty. These two issues cause major problems for new users (as the system has difficulty making initial recommendations) and previous existing users (recommending new items when available/possible instead of the same items every time).

This research aims to deal with these issues and apply non-redundant multi-level and cross-level association rules to recommender systems.

The paper is organized as follows. Section 2 presents the motivation for our work. Related work is briefly presented and detailed in Section 3. Our research problems and questions are listed in Section 4. Section 5 outlines the approach our research is taking. Current and expected future contributions are given in Section 6. Lastly, Section 7 concludes the paper.

2. Motivation & Rationale

When it comes to association rule mining a lot of effort has been focused on efficiency (speed) and effectiveness (number). Less effort has been focused on quality. Often when using association rule mining many rules can be discovered or derived and overwhelm the user. But more importantly some of these rules could be redundant and bring no new knowledge. Some effort has been directed at dealing with redundant rules in flat datasets [13,14,9]. However datasets can have a hierarchy/taxonomy or multiple concept levels and thus redundancy in these

978-0-7695-3496-1/08 $25.00 © 2008 IEEE
DOI 10.1109/WIIAT.2008.39

datasets need to be focused on. This issue is one of the aspects of this research.

Another issue when it comes to quality association rules is not just redundant rules, but rather how the quality of a rule is measured to determine if it is useful, interesting, important etc. How the quality of an association rule is determined is important, but there is no formal definition of quality and/or interestingness [2]. Currently there is a collection of different measures available which is partly due to the traditional methods of support and confidence being considered insufficient [7]. Because of this and the fact that most of these measures have focused on association rules derived from flat datasets, we believe a better measure for determining and measuring the quality or interestingness of multi-level and cross-level association rules from datasets with multiple concept levels is needed.

Recommender systems are a tool designed to help users by giving information recommendations according to each user's information needs and have found uses in many environments. There are two major approaches used in recommender systems; content based approach and collaborative approach. Both approaches rely on data from users (such as ratings of items) in order to make recommendations. This leads to two problems, cold start (insufficient initial information) and recommendation novelty (the ability for a system to broaden a users' interest over time or recommend new items and not just the same ones). We believe overcoming these issues is important in order to improve recommender systems. Taxonomy/hierarchy information from a dataset can provide a resource to improve upon these issues. Currently we believe this resource is not utilized well enough and our research plans to make use of it through association rule mining to improve upon recommender systems.

3. Related Work

Since its introduction in [1], association rule mining has become both an important and widely used data mining technique. The aim of this technique is to extract frequent patterns, interesting correlations and associations amongst sets of items in large transactional databases.

Much work in the field of association rule mining has focused on finding more and more efficient ways to discover all of the rules possible. This has meant less work has focused on the issue of the quality of the discovered association rules.

Currently the approach being taken is to determine which rules are redundant and remove them, thus reducing the number of rules a user has to deal with while not reducing the information content [8,12]. These approaches show a lot of promise and work by Xu & Li [12] shows that a reduction of over 80% for exact basis rules can be achieved. More recently this work was extended to also include removing redundancy in approximate basis rules [13]. These works have only focused on datasets where all items are at the same concept level. Thus they do not consider redundancy that can occur when there is a hierarchy among items.

Work has been done in adapting approaches originally made for single level datasets into techniques usable on multi-level datasets. Han & Fu's work [3] shows one of the earliest approaches proposed to find frequent itemsets in multi-level datasets and later was revisited [4]. This work primarily focused on finding frequent itemsets at each of the levels in the dataset and did not focus heavily on cross-level itemsets (those itemsets that are composed of items from two or more different levels). In fact the cross-level ideas were an addition to the work being proposed. Further work proposed an approach which included finding cross-level frequent itemsets [11].

However, even with all this work, the focus has been on finding the frequent itemsets efficiently and the issue of quality and/or redundancy in single level datasets. Some brief work by Han & Fu [3] discusses removing rules which are hierarchically redundant, but it relies on the user giving an expected confidence variation margin to determine redundancy. Thus there appears to be a void in dealing with hierarchical redundancy in association rules derived from multi-level datasets. One of the objectives of this research program is to fill this void and solve the issue of hierarchical redundancy.

Because of the overwhelming information overload problem, high quality recommender systems are in big demand. High quality recommender systems will benefit users as more accurate systems can recommend more highly relevant items, including new items and also recommend different items over time based on changes to the dataset(s) and changes to what the user has already viewed.

Recommender systems provide methods of finding important and relevant information from large quantities of data with minimal human effort. Previous work has shown that association rule mining techniques can be applied to recommender systems to improve their accuracy [5,6]. However, the current existing work still fails to solve the cold start problem since they only utilize the associations between users and items based on previous users' rating data. This project seeks to improve on the cold-start and novelty

issues through new approaches that utilize multi-level and cross-level association rules from taxonomy data.

4. Research Problems & Questions

Ultimately the aim of this research is to contribute to the field of data mining and in particular to association rule mining and recommender systems. Thus to achieve this clear and well defined research problems and questions are needed. By solving these it is possible to contribute in a meaningful manner.

For our research there are three main problems or questions that we will focus on and attempt to solve. These problems/questions are:

1. To define what redundancy in multi-level datasets is and effectively and efficiently discover non-redundant multi-level and cross-level association rules from datasets with a hierarchy/taxonomy.
2. Design an approach to comprehensively determine and evaluate the quality of an association rule or set of association rules derived from a dataset with multiple concept levels.
3. Developing an approach whereby the discovered non-redundant multi-level and cross-level association rules are utilized by a recommender system to solve the cold start and/or novelty problems.

5. Research Approach

The proposed research has the following aspects; exploration, observation, experimentation, prediction & confirmation and description & explanation. The design, methodology and approach(es) chosen have to take into account these aspects and utilise them effective in manner that will be of benefit.

Based on our proposed research, three methodologies have been chosen and include the scientific research method, action research and experimental research as it is believed that the combination of these methodologies best fits the approach needed for our research. This allows the strengths of these three approaches to be combined.

It is possible to conduct experiments in conjunction with the scientific method and often when so done, the experiment can then provide the best conclusion about the hypothesis under test. For this research, the type of experiments being undertaken will be of the controlled type. A controlled experiment compares the outcomes/results from an experimental sample with the outcomes/results from the control sample. The only difference between these two samples is the factor being tested / assessed. Both samples are tested using exactly the same experiment so that a reliable and valid comparison can be made. The results of the experimentation will then be used to support or refute the hypotheses/predictions that were developed using the scientific approach and help to refine our work until a successful implementation or result is achieved. The results from the experiments also need to be reproducible; so that other researchers can conduct the same work, get the same results to help confirm or refute the outcomes and improve upon the outcomes achieved in this research program.

The design & approach being undertaken will also include attributes from both quantitative and qualitative research and thus our research will be a combination of several different aspects brought together.

6. Research Contributions

Although this research is still in progress, at this stage significant steps forward have been achieved. As previous detailed, the first problem that this research is attempting to solve is removing redundant association rules from datasets with multiple concept levels. Towards solving this problem we have successfully defined what we call hierarchical redundancy for both exact association rules and approximate association rules [9,10].

We have defined hierarchical redundancy in exact basis association rules as:

***Definition 1*:** Let $R_1 = X_1 \Rightarrow Y$ and $R_2 = X_2 \Rightarrow Y$ be two exact association rules, with exactly the same itemset Y as the consequent. Rule R_1 is redundant to rule R_2 if (1) the itemset X_1 is made up of items where at least one item in X_1 is descendant from the items in X_2 and (2) the itemset X_2 is entirely made up of items where at least one item in X_2 is an ancestor of the items in X_1 and (3) the other non-ancestor items in X_2 are all present in itemset X_1.

Likewise, we have also defined hierarchical redundancy in approximate basis association rules as:

***Definition 2*:** As per Definition 1 with the extra condition (4) the confidence of R_1 (C_1) is less than or equal to the confidence of R_2 (C_2).

From these two definitions we have built upon the MinMaxExact and MinMaxApprox algorithms by Pasquier et. al. [8] and the RelaibleExactRule and ReliableApproxRule algorithms by Xu et. al. [12,13] to build algorithms that will also remove hierarchically redundant exact and approximate association rules [9,10].

Tables 1 & 2 show some of the early results obtained in our work when we apply our definitions (1 & 2) during rule extraction. As can be seen, our

extension (shown as 'with HRR') reduces the size from between 9 & 25% for the exact basis rule set and between 12 & 24% for the approximate basis rule set.

Table 1. Results for hierarchical redundancy removal in exact basis association rules.

	Exact Basis						Exact Rules
Data set	MME	MME with HRR	%	RER	RER with HRR	%	
T3	174	134	23	113	89	21	736
T4	577	429	25	383	305	20	1584
T5	450	405	10	315	287	9	725
T6	725	602	17	91	80	12	725

Table 2. Results for hierarchical redundancy removal in approximate basis association rules.

	Approximate Basis						Approx Rules
Data set	MMA	MMA with HRR	%	RAB	RAB with HRR	%	
T3	700	587	16	398	347	12	1447
T4	2546	2085	18	1608	1387	13	4332
T5	6427	4844	24	3415	2970	13	NA

Also, by the completion of this research we hope to have been able to make the following further contributions:

1. A measure/approach for determining and evaluating the quality of association rules and/or association rule sets (including both multi-level and cross-level rules) derived from datasets with multiple concept levels.
2. An approach and implementation of a recommender system that utilizes and takes advantage of the non-redundant multi-level and cross-level association rules derived from a dataset with a hierarchy/taxonomy to overcome the cold start and novelty problems when making recommendations to users.

7. Conclusions & Future Work

In this paper we have presented a brief overview of research that we are currently undertaking. Progress has been made with some contributions to the research field already achieved.

Future work from this research can include continuing to investigate redundancy in association rules to reduce the size of the rule set, along with working upon the quality measure proposed to allow for better measuring of a rule's quality. By improving the quality of these rules, improvements can be made to recommender systems that utilize them. And by continuing to work on recommender systems that use association rules, the cold start and novelty problems may be overcome.

8. References

[1] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases," in *ACM SIGMOD International Conference on Management of Data (SIGMOD'93)*, Washington D.C., USA, 1993, pp. 207-216.

[2] L. Geng and H. J. Hamilton, "Interestingness Measures for Data Mining: A Survey," *ACM Computing Surveys (C SUR),* vol. 38, 2006.

[3] J. Han and Y. Fu, "Discovery of Multiple-Level Association Rules from Large Databases," in *21st International Conference on Very Large Databases*, Zurich, Switzerland, 1995, pp. 420-431.

[4] J. Han and Y. Fu, "Mining Multiple-Level Association Rules in Large Databases," *IEEE Transactions on Knowledge and Data Engineering,* vol. 11, pp. 798 - 805, Sep/Oct 1999.

[5] C. Kim and J. Kim, "A Recommendation Algorithm Using Multi-Level Association Rules," in *IEEE/WIC International Conference on Web Intelligence*, 2003, pp. 524-527.

[6] W. Lin, S. A. Alvarez, and C. Ruiz, "Efficient Adaptive-Support Association Rule Mining for Recommender Systems," *Data Mining and Knowledge Discovery,* vol. 6, pp. 83-105, Jan 2002.

[7] K. McGarry, "A Survey of Interestingness Measures for Knowledge Discovery," *The Knowledge Engineering Review,* vol. 20, pp. 39-61, Mar 2005.

[8] N. Pasquier, R. Taouil, Y. Bastide, G. Stumme, and L. Lakhal, "Generating a Condensed Representation for Association Rules," *Journal of Intelligent Information Systems,* vol. 24, pp. 29-60, 2005.

[9] G. Shaw, Y. Xu, and S. Geva, "Eliminating Redundant Association Rules in Multi-level Datasets," in *4th International Conference on Data Mining (DMIN'08)*, Las Vegas, USA, 2008, p. To appear.

[10] G. Shaw, Y. Xu, and S. Geva, "Deriving Non-Redundant Approximate Association Rules from Hierarchical Datasets," in *ACM 17th Conference on Information and Knowledge Management*, Napa Valley, USA, 2008, p. To appear.

[11] R. S. Thakur, R. C. Jain, and K. R. Pardasani, "Mining Level-Crossing Association Rules from Large Databases," *Journal of Computer Science,* vol. 2, pp. 76-81, 2006.

[12] Y. Xu and Y. Li, "Generating Concise Association Rules," in *16th ACM Conference on Conference on Information and Knowledge Management (CIKM'07)*, Lisbon, Portugal, 2007, pp. 781-790.

[13] Y. Xu, Y. Li, and G. Shaw, "Concise Representations for Approximate Association Rules," in *IEEE International Conference on Systems, Man & Cybernetics (SMC'08)* Singapore: IEEE, 2008, p. To appear.

[14] Q. Zhao and S. S. Bhowmick, "Association Rule Mining: A Survey," Nanyang Technological University, Singapore, 2003.